Montana State Library
Spatial Data Transfer Design

DRAFT prepared for GeoMAPP, December 6, 2011
(last edited: Papineau)

Introduction

The Montana State Library (MSL) was integrated into the Geospatial Multistate Archive and Preservation Partnership (GeoMAPP) as a full partner in February 2011 after a little over a year as a GeoMAPP informational partner. MSL has served as the state GIS clearinghouse for almost two decades. In this capacity, MSL manages a large spatial data collection and makes GIS data available via web applications, web map services, and as downloadable data. For years MSL has maintained a GIS data list, and in 2008 MSL launched the Montana GIS Portal based on the ESRI GeoPortal Toolkit.

MSL has long been recognized as the "archives" for GIS data, though informally executed. Prior to joining GeoMAPP, MSL's process to archive GIS data was to simply not throw any data away. As MSL approached the GeoMAPP project, MSL chose to take a "library collection development policy" approach to managing a GIS data archive rather than a "records management" approach common in government document management.

MSL manages its own internal data center supporting the GIS clearinghouse and the informal GIS archive. Storage is comprised of an SQL database for managing active datasets and a file structure created on a Storage Area Network (SAN) for Dark Archive storage. All data storage is backed up meeting Montana State standards.

Note: This data transfer demonstration effort focuses only on archiving spatial data stored and served by MSL. It does not cover archiving MSL's electronic maps, paper maps, map project files, text documents, or web pages.

Planning for Incorporating Preservation Practices

MSL staff members spent a significant portion of 2011 reviewing GeoMAPP documentation and envisioning how to fit archiving into Montana's geospatial clearinghouse workflows. The GeoMAPP data transfer best practice and design documentation taught MSL about practices and challenges applicable to archiving spatial data. It also spawned an objective critique of MSL's existing data storage, data access tools, and management processes in light of what's needed to consistently and professionally archive spatial data.

One early task involved clarifying four important terms and how they would define and support a revised data management process that includes archiving as MSL moves forward:

> *Data Collection*—data MSL stores and preserves that fits the draft Collection Development Policy. The collection may include multiple copies of data in different forms stored in the Dark Archive as well as the Accessible Archive and the Active Store.
>
> *Dark Archive*—a physical location on MSL's SAN, storing preservation copies of data MSL accepts into the data collection. The Dark Archive serves as the master copy of MSL spatial data accessible online. Data that is too large for an online offering will be manually extracted from the Dark Archive to meet patron requests. The Dark Archive will ultimately store preservation copies of electronic maps and GIS map projects. The Dark Archive will not permanently store spatial data that is archived by other organizations such as the National Agricultural Imagery Program (NAIP) imagery and wetlands data.
>
> *Accessible Archive*—a physical storage location offering superseded data made easily accessible to patrons via the internet. This is older data that is not of interest to most patrons, but is valuable to many, justifying a dedicated volume and access tools for data download (i.e. superseded Cadastral, Land Cover, etc.). The storage and service vendor has yet to be determined.

Active Store—a group of MSL SAN storage locations holding the most active data (primarily the most current data) made available in different forms for online, self-service access. Different storage locations will serve data for internal use in an SDE database, as downloadable files, as web mapping services, and serving online mapping applications. Data in Active Store locations is data that will be used most frequently by library patrons (internal and external).

Clearinghouse—a group of data discovery tools and resources, including the Montana GIS Portal, web mapping applications, web map services, webpages offering data download, as well as staff time for manual packaging of large datasets for patrons.

The next effort undertaken as MSL moved toward the data transfer demonstration involved envisioning a new, streamlined data management process that incorporates formal archiving practices and concepts. To start this process, MSL documented existing data repositories and file flow processes. This enabled the modeling of natural spatial data clearinghouse workflows that could be mimicked in a new data management system. With this draft data management process in hand (Appendix A), MSL wrote initial technical requirements for a new data and metadata management system to automate as much of this data management work as possible. The data and metadata management system is currently in development. It will assist with spatial data and metadata ingest, metadata updates, metadata publishing, data management, and data integrity auditing.

Considering the abbreviated period of time involved with GeoMAPP (less than one year), it was not possible to create, test, and deploy a new management system prior to completing the GeoMAPP data transfer demonstration. Instead, MSL created a prototype environment that would mimic the new management system under development. MSL created a Dark Archive file structure on the SAN that reflects the pattern by which data is accessioned into MSL's collection. (Figure 1, right). MSL also tested and chose to use the Library of Congress's tool "Bagger"[1] for data transferring to the Dark Archive and validating. This GUI-driven tool (Figure 1, left) automates many archival processes, including data packaging, creating a manifest of the contents, creating a checksum, and validating file integrity.

[1] Bagger available via http://sourceforge.net/projects/loc-xferutils/files/loc-bagger

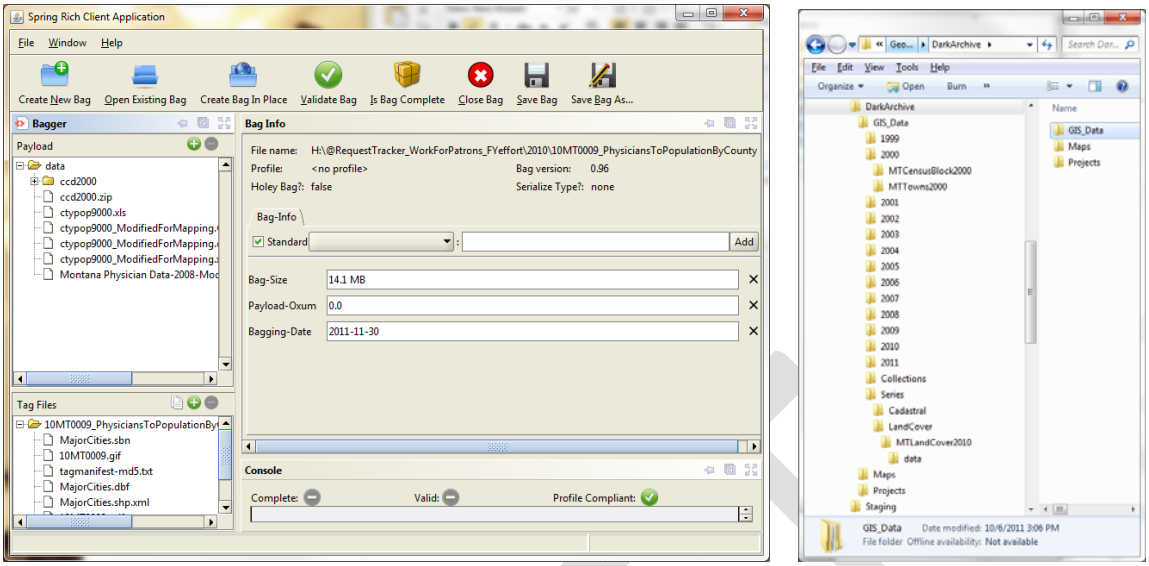


Figure 1: Bagger User Interface (left) and MSL Spatial Data Dark Archive File Structure (right)

MSL chose to archive data in its existing formats rather than define an archival format because each data format offers unique functionality. Though MSL realizes that shapefiles are usable in many GIS software programs, the team did not want to remove functionality inherent in those file types by converting all data to shapefiles as an archive standard type. The datasets selected for transfer included:

- Series Data—Landcover file geodatabase (raster data, 100MB)
- Stand-Alone Data—Montana Towns shapefile (.08MB)
- Stand-Alone Data—Montana Census Blocks shapefile (57MB)

MSL established a download folder and file naming convention in the form of <*extent*><*theme*><*time period*> (i.e. MTcensusblocks2010) and created a mock-up data management spreadsheet to mimic the proposed data management system's database. This mockup spreadsheet had the following fields defined:

FGDC metadata—Title, Time Period, Originator, Publisher, Other Online Location, Larger Work Citation

New Archival metadata—Date Archived, Zipped Megabytes, Checksum, Last Archive Review Date, Accessible Archive Online Location

New Administrative metadata—Clearinghouse Online Location, File Location, Data Format, Data Format Version, SIP Metadata URL/network location, DIP Metadata Blob, Source Data (if derived), Derivatives, Compression, Status

Note: Complete user metadata records, including any additional archive metadata, will be stored in blob fields in the Montana GIS Portal. Full metadata records stored in blob fields enables data discovery by patrons. The metadata recorded in the data management system database aids in data management and archiving, not patron data discovery.

Data Transfer Demonstration

With these initial planning tasks completed, MSL proceeded with the actual data transfer demonstration to test these ideas and record any problems encountered. For the demonstration, MSL worked through the following procedure and modified it with each use. The last dataset that was transferred used the following process, representing MSL's draft data transfer process to date:

1. Select data for transfer that meets the requirements of the draft Collection Development Policy and place it in the Dark Archive staging folder.

2. Run a virus check on the data using ESET NOD32. Even though NOD32 is always checking for viruses on network storage devices, it is best to execute this virus check specifically on the file before placing the data in the Dark Archive.

3. Confirm that the metadata is valid—read metadata add descriptive information if necessary.

4. Consider preserving the original file name or rename it to meet the file naming convention.

5. Pre-populate some metadata fields in the data management spreadsheet, pulling information from the data's FGDC metadata record (Title, Time Period, Originator, Publisher, Other Online Location, Larger Work Citation).

6. Using Windows Explorer and Bagger, package the data for transfer to the Dark Archive:

    a. In Windows Explorer, create a compressed zip folder using the file naming convention.

    b. Place the dataset contents in the zip folder. The zip folder includes the data and a user metadata record.

    c. With Bagger, create new bag. Set parameters as version 9.6 with no profile.

d. Add files using Bagger's green plus icon (Figure 1) and navigate to and select the files for bagging then click Open.

e. Save the bag: File>Save Bag As.

f. In the Save In text box, use the browse button to choose the location in the archives where the bag will be placed. In this dialog box, name the location using the file naming convention.

g. Accept defaults in the Save In dialog box, but ensure that the data will not be zipped by Bagger. The transferred bag will include: the data and user metadata in a Data folder, plus other Bagger-generated files such as a checksum and manifest as shown in Figure 2.

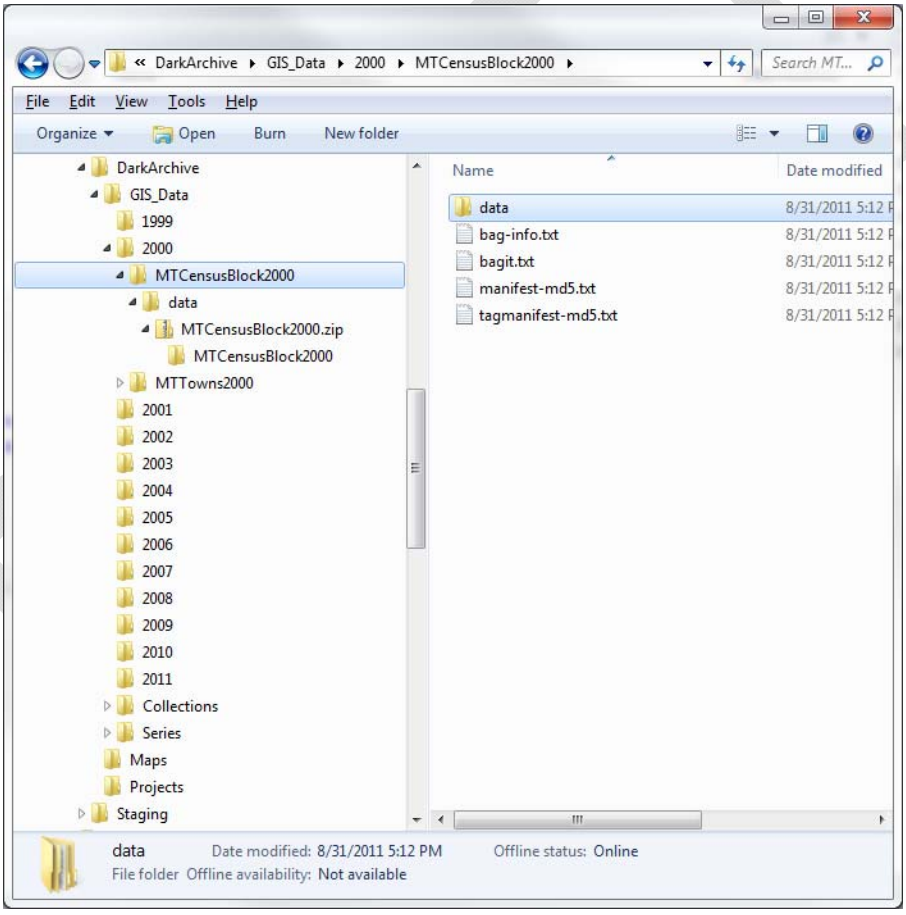


Figure 2: Bagger Generated Files with Data Transfer

h. Validate the bag using Bagger's Validate Bag button. Check that the bag is complete using the Is Bag Complete button (Figure 1).

i. Close the bag.

7. Enter archival and administrative metadata in the management system spreadsheet, including recording the zip file's checksum.

8. Delete dataset from the staging area.

Data Transfer Evaluation

Overall, the data transfer demonstration task met and even exceeded expectations. The data transfer rates did not vary from what MSL has encountered previously as the state's GIS Clearinghouse; therefore it did not alter MSL's standard approach of transferring large datasets after work hours to avoid slowing down the network. The data transfer rates for the demonstration data were as follows:

- Landcover geodatabase zip file (100MB): 30 seconds
- Montana Census Blocks shapefile zip file (57MB): 7 seconds
- Montana Towns 2000 shapefile zip file (84KB): 1 second

MSL took a different approach with defining the Dark Archive data storage structure from that recommended in the GeoMAPP Data Transfer Best Practice. GeoMAPP suggests that data in an archive be categorized into one of nineteen ISO 19115 Topic Categories and stored in folders named for these categories. MSL considers ISO topic categories critical to how data is discovered by patrons. However, using them to guide staff with data storage placement presents difficulties because data often can be classified as meeting more than one ISO topic category. Data selection of one ISO category for a dataset is then subject to staff member preferences, which may vary within the team. MSL chose instead to ignore ISO categories when designing the Dark Archive storage system and chose to simply store the data by the time period of content and then by the data set theme. The patron data discovery system is the place where ISO categories will be leveraged. Parent-level collection and series metadata records will report the association between datasets stored in different time period folders.

MSL chose to zip the data before storing in the Dark Archive to both save space in the archive and to permit easier downloading by patrons. Also, bagging one zip file instead of loose files produces one checksum, which may streamline data management and data integrity checking in the workflow. MSL chose against using the Bagger zip functionality because the resulting data package made available to patrons had an excessively deep file structure, burying the data in multiple levels of folders.

The nature of how MSL approached the data transfer demonstration task was to learn from each transfer and strengthen the process. MSL revised the procedure between

each run, continuing to test the procedure each time. The process recorded in this document represents the draft procedure to date, though there are several areas that need further testing and consideration.

Though MSL will be using collection and series metadata records, the Montana Metadata Work Group (led by MSL) is wading through some associated challenges these records present. These "parent" metadata records will be useful to MSL staff when performing data management tasks, and for streamlining and simplifying patron search results when a data query returns many dataset records that are all related.

For existing datasets in the collection, MSL chose to use existing file names, which are based on an older, MSL file naming convention. The significant effort required to change these in a large collection (and associated discovery tools) is too great with limited resources. However, existing datasets that are en route to the Dark Archive will be packaged into zip files that do use the new naming convention. Newly-accessioned datasets will use the file naming convention wherever naming is required. For the existing datasets using the old naming convention, MSL may devise a tool that will apply the file naming convention to data files as they're downloaded to patrons.

The draft data management process (Appendix A) suggests that Submission Information Packages (SIPs) are defined as datasets that may be modified by MSL staff before archiving to make the data more usable to library patrons. There is some question whether the exact original submission will also be archived. Essentially, MSL needs to decide what a spatial data SIP is or if it will vary from dataset to dataset.

Next Steps

Further work needs to be completed to more formally define the MSL Spatial Data Collection Development Policy. Emphasis is placed on Montana GIS Clearinghouse data which has a statewide focus and MSDI data which incorporates both local and state data.

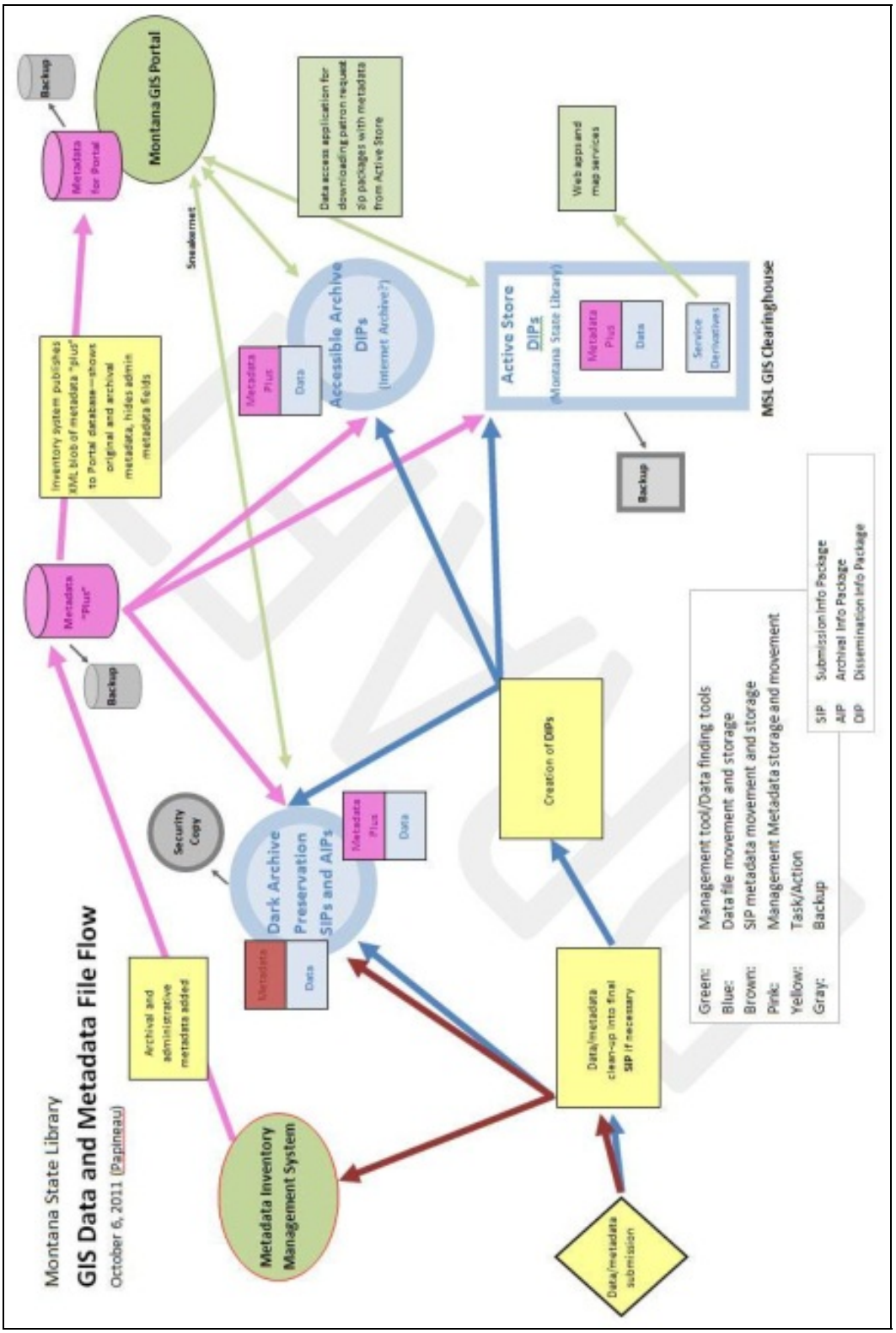
MSL has draft system requirements for an initial release of a data management system in development (Appendix B). This system will be used to continue the on-going process to inventory and archive the extensive MSL GIS data collection. Once MSL works through some of the issues identified in the data transfer evaluation, MSL will update these system technical requirements. In this GeoMAPP deliverable, the system requirements represent what is known to date. Additionally, MSL staff will work with all Montana Spatial Data Infrastructure (MSDI) theme stewards to develop archiving plans for each theme.

More research will also be conducted on ways to improve discovery and delivery of archived GIS data. An initial review of the Internet Archive's http://archive.org site shows that a handful of shapefiles are available for download. MSL plans to test distribution of GIS data via this site in conjunction with the MSL State Publications program.

MSL also plans to modify the State of Montana's GIS metadata technical specifications document to include required archival fields. This change will improve discoverability of archived spatial data through the Montana GIS Portal. Because the process to update this document will require input from the Montana Association of Geographic Information Professionals and the State of Montana GIS Managers Forum, it will provide a good opportunity for MSL staff to educate Montana's GIS community about the criticality of metadata to the archiving process.

Montana is currently exploring different funding models to fund GIS data development activities including on-going development of the MSDI. An initial funding proposal includes an annual budget for a GIS archives program. This funding need will continue to be pursued as part of the larger funding discussion to be taken to the 2013 Montana Legislature.

Appendix A: *Draft* Data Management File Flow



Appendix B: *Draft* Spatial Data Management System Requirements

The system shall be able to:

1. Ingest a set of certain metadata fields into the system and validate against requirements.
2. Push SIP metadata from bagged file into a Blob field (SIP Blob).
3. Ingest data packages to the archives.
4. Generate series and collection IDs.
5. Generate unique IDs for each dataset that is recorded in its metadata.
6. Move data packages from one location to another.
7. Generate a checksum upon data ingest and record that checksum in a database field.
8. Auto-populate the location of the data destination upon transfer.
9. Auto-populate other admin metadata fields that you can, such as date uploaded.
10. Permit certain automatically-populated fields to be human edited after populating and then not overridden by the system.
11. Be able to distinguish between original metadata uploaded and metadata fields edited and used for administration and archiving.
12. Be extensible for future growth (adding metadata fields and when possible asking the system to populate those fields by surveying the data and metadata holdings; modifications to any system components such as a database, the GUI, report formats, etc.).
13. Permit migration of the contents of the system to any new system as technology changes.
14. Permit removing of test metadata records or mistakes with the system. Button "Remove" but it marks that record as removed (not deleting actual record)
15. Perform regular audits/validations of data integrity.
16. Send notification if an audit/validation check error occurs.
17. Send notification when a dataset/collection/series is due for reappraisal.
18. Record actions in an audit trail.
19. Configure access and permissions.

20. Be able to recognize that newly-ingested data is part of an existing collection or series leveraging a series or collection ID.

21. Be able to identify parent/child relationships among data and among metadata records

22. Automatically push select XML metadata out into a blob field in a separate database for use with the Montana GIS Portal.

23. Create Administrative metadata fields in the system that are populated manually, including a Notes or Comments field and an Other field, and the ability to attach documents

Ideally, the system will be able to:

1. Harvest and upload data from locations outside of the Montana State Library's network.

2. Generate reports based on predefined and ad hoc queries.

3. Be able to view/edit/move records and data in bulk via queries against the metadata (see next entry).

4. The system shall be able to respond to queries that answer the following questions:

    - What data needs to be reviewed related to its presence in the Dark Archive?
    - What data needs to be updated?
    - What data is used in <this> web application?
    - What data is used in web map services?
    - What data covers <this> part of Montana?
    - What data is downloadable?
    - Which data is framework data?
    - What data is stored in the Accessible Archive?
    - What data is stored in the Active Store?
    - What data is in Purgatory?
    - What data take up a lot of space?
    - What data was created by <this> agency?